Université de Caen Normandie

IUT Grand Ouest Normandie

Département Science des Données

Analyse de la série chronologique représentant le nombre de nuitées dans l'hôtellerie en Normandie

Yolan DENESLES - Lou-Anne THOMAS - Oscar JOSEPH–GENESLAY

11 janvier 2024

Résumé

L'Institut national de la statistique et des études économiques (INSEE) met à disposition du grand public des séries chronologiques portant sur diverses thématiques. Dans le cadre de cette étude, notre attention se porte tout particulièrement sur la série concernant le nombre de nuitées dans l'hôtellerie en Normandie, identifiée par le code 010598624. Ces données couvrent la période de janvier 2011 à septembre 2023 et exprimées en milliers.

Table des matières

1 Présentation des données

Les données proviennent de l'INSEE (code 010598624) représentant le nombre de nuitées dans l'hôtellerie en Normandie sur la période janvier 2011 à septembre 2023. Les données sont exprimées en milliers. Nous avons récolté ces données et nous les avons insérées dans un dataset. Voici un extrait des données :

	Date	Value
1	2011-01-01	344
2	2011-02-01	416
3	2011-03-01	504
4	2011-04-01	722
5	2011-05-01	726
6	2011-06-01	855
7	2011-07-01	881
8	2011-08-01	961
9	2011-09-01	771
10	2011-10-01	691

Nous avons dans celui-ci deux colonnes :
— **Value** : représentant le nombre de nuitées sur la date indiquée.
— **Date** : représentant la date sous le format (Année-Mois-Jour) de l'enregistrement de la valeur.

2 Analyse exploratoire

2.1 Visualisation de la chronique

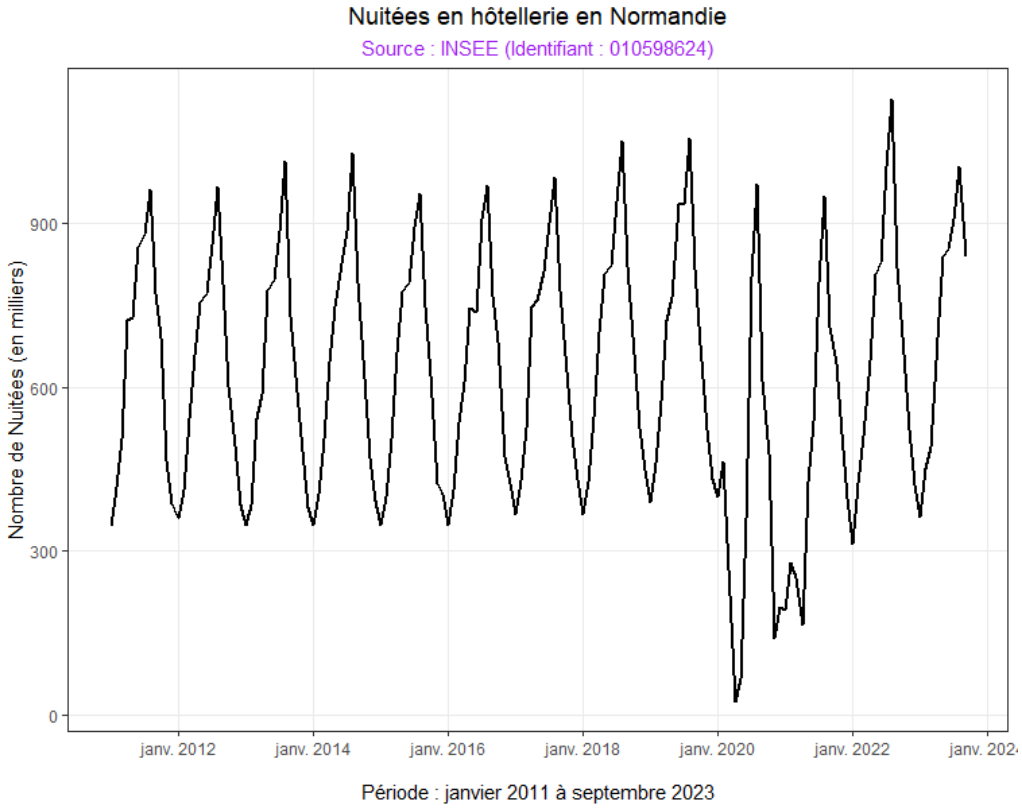
Le chronogramme ci-dessous, issu des données de l'INSEE (identifiant : 01058624), représente l'évolution du nombre de nuitées en hôtellerie en Normandie, en milliers d'unités, sur la période de janvier 2011 à septembre 2023.

On observe que la série temporelle présente trois caractéristiques principales :

— Une **saisonnalité annuelle**, avec des fluctuations régulières du nombre de nuitées selon les mois de l'année. On distingue des pics de fréquentation en juillet-août et en décembre, et des creux en janvier-février et en septembre-octobre.

— Une **tendance générale**, qui indique la direction et le rythme de l'évolution du nombre de nuitées sur le long terme. On remarque que la tendance est relativement stable jusqu'à janvier 2020 où l'on observe une rupture de la tendance, avec une chute brutale du nombre de nuitées, suivie d'une remontée spectaculaire sur juillet-août 2020. Après cette période de forte volatilité, la tendance semble se stabiliser. De plus, on observe une valeur atypique en avril 2020.

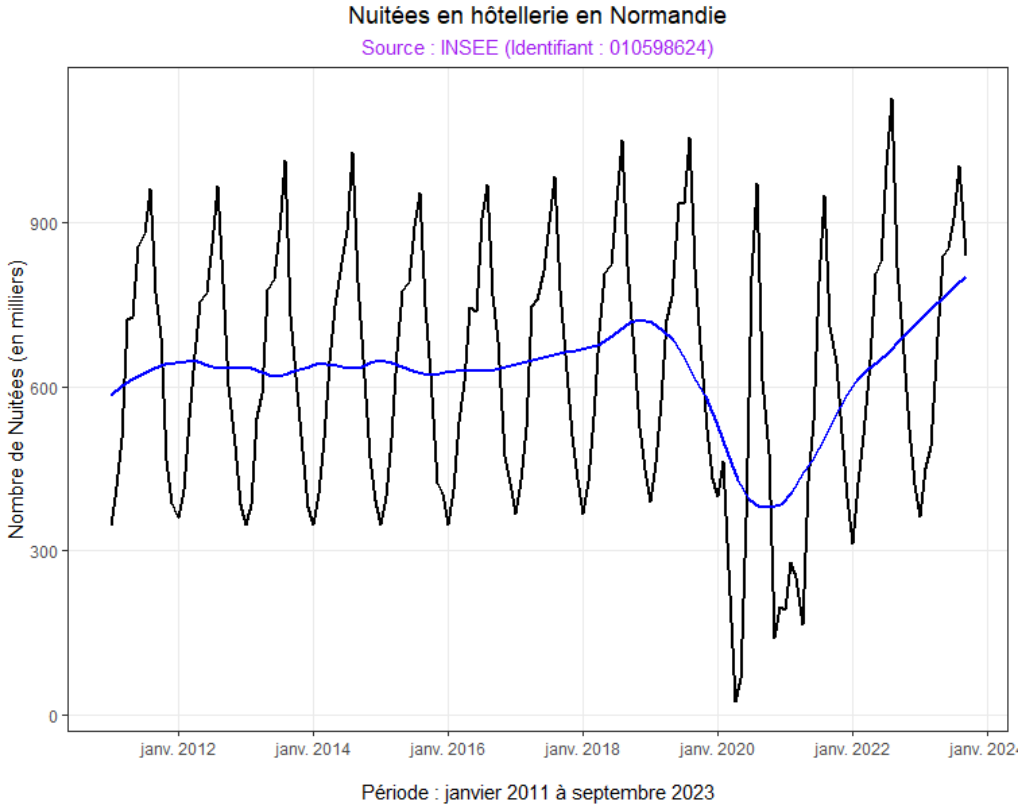
— Le chronogramme forme un **schéma additif**, en effet les enveloppes supérieures et inférieures semblent parallèles.



2.2 Visualisation de la chronique avec une tendance lissée

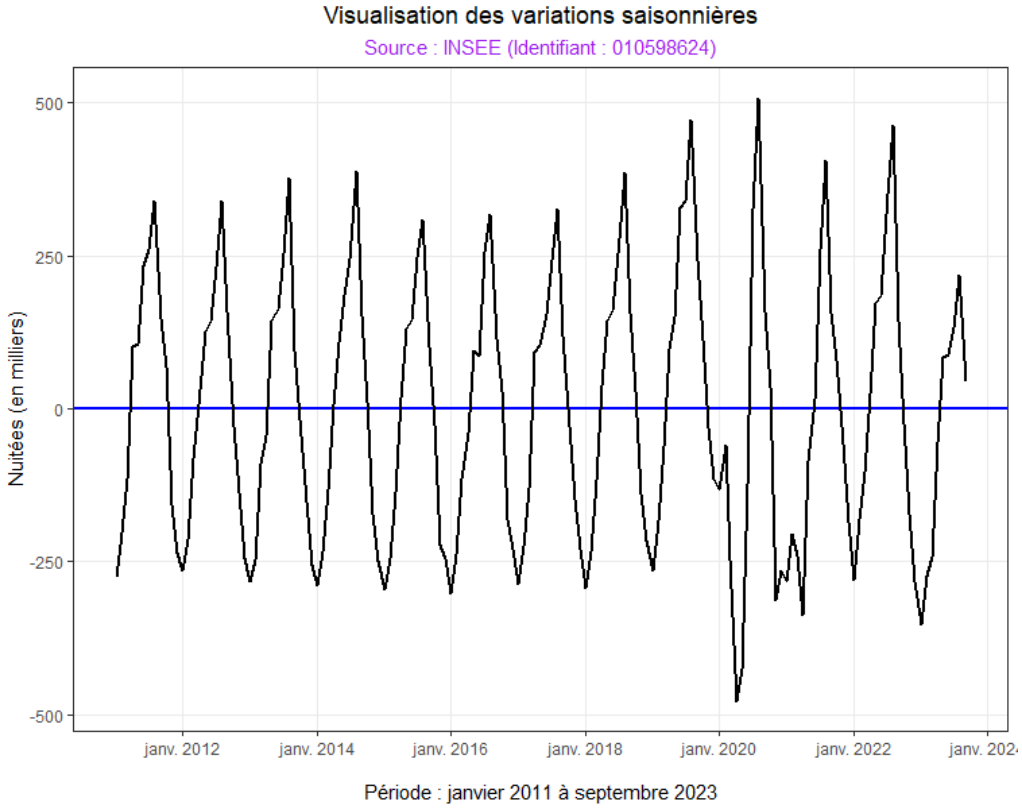
Pour mieux appréhender l'évolution de la tendance, nous avons utilisé une méthode non-paramétrique de lissage. Nous avons obtenu la courbe de tendance lissée, représentée en bleu sur le graphique ci-dessous.

Ainsi, la tendance générale ne semble pas être linéaire, car notre densité lissée ne semble pas suivre une droite. Cependant, nous visualisons plus clairement deux points de ruptures, l'un en décembre 2018 et le second en octobre 2020. Sur la première période (janv. 2011 à déc. 2018) la tendance semble suivre une droite linéaire très légèrement positive. Sur la deuxième période (dec.2018 à oct. 2020), on observe une droite linéaire négative. En effet, la tendance du nombre de nuitées connaît une chute brutale suite aux confinements. Sur la troisième période (oct. 2020 à sept. 2023), le nombre de nuitées semble se stabiliser à la hausse avec une légère remontée en juillet-août 2021 au-delà des observations faites avant la rupture. Cette tendance lissée prend la forme d'une droite linéaire positive.



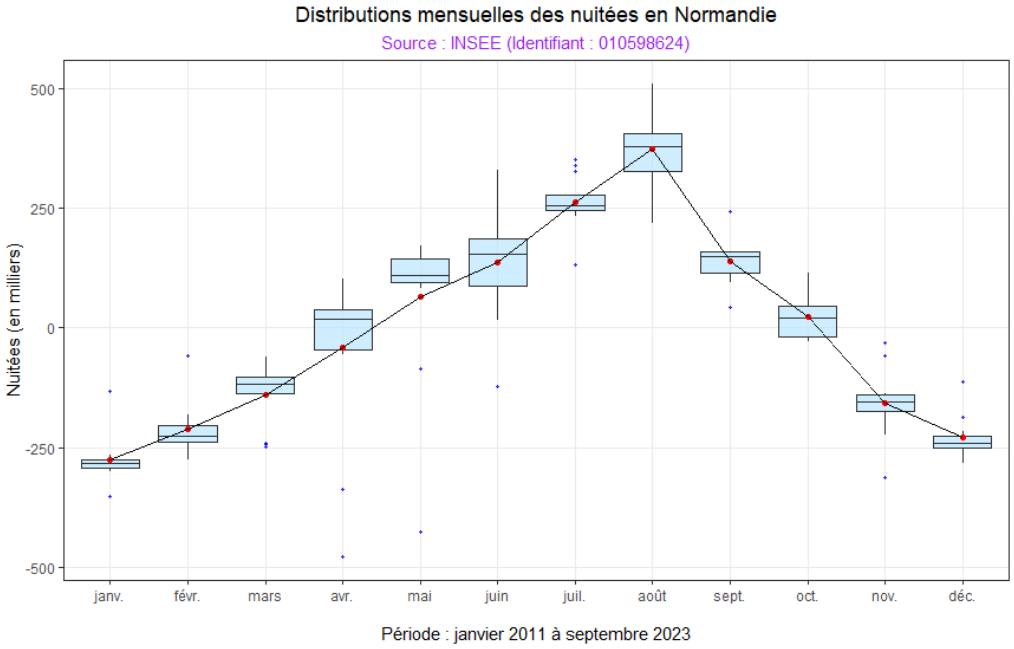
2.3 Visualisation des variations saisonnières sous forme de chronique

Ce graphique représente les variations saisonnières sous la forme d'une chronique. L'hypothèse d'une composante saisonnière ne peut pas être remise en question parce que l'on visualise bien la structure d'un motif qui se répète. On remarque toujours notre valeur atypique en avril 2020. De plus, on observe que les variations saisonnières semblent se compenser autour de 0. En effet, les variations saisonnières sont distribuées de façon symétrique avec des fluctuations équivalentes tant au-dessus qu'en dessous de la ligne représentant la valeur 0. Un point atypique, identifié précédemment en avril 2020, se démarque par une déviation significative par rapport au schéma habituel des variations saisonnières. Cela coïncide avec la période de confinement due à la pandémie de Covid-19, marquant ainsi un événement exceptionnel dans l'analyse.



2.4 Distribution mensuelle des variations saisonnières

Ce graphique à boîtes à moustaches représente la distribution mensuelle des variations des nuitées en Normandie par mois. On observe que les mois de juin, juillet, août et septembre sont les mois les plus favorables aux nuitées dans les hôtels. Ce phénomène pourrait être dû à la présence des vacances d'été qui attire massivement les clients. À l'inverse, les mois les plus défavorables sont les mois de janvier, février, mars, novembre et décembre. Les mois d'avril et octobre ne présentent pratiquement pas de fluctuations. On peut également visualiser la variabilité d'un mois d'une année à l'autre d'après l'amplitude des boxplots. Une amplitude plus importante suggère une plus grande dispersion des données, indiquant ainsi des variations significatives dans le nombre de nuitées d'un mois d'une année à l'autre. Ainsi, les mois d'avril, de juin et d'août, sont plus susceptibles de subir des variations notables du nombre de nuitées d'une année à l'autre.



3 Modélisations

3.1 Modèle de régression linéaire par morceaux

On se propose maintenant d'ajuster cette estimation de la tendance à l'aide d'une régression linéaire simple par morceaux. En effet, une fois les points de rupture trouvés, on peut diviser la série temporelle en sous-périodes, et ajuster un modèle de régression linéaire simple sur chaque sous-période. Pour rappel, nos points de ruptures sont les suivants : t1 correspond au mois de décembre 2018 et t2 au mois d'octobre 2020.

Notre modèle de régression linéaire par morceaux suit ainsi l'équation suivante :

$$m(t) = \beta_0 + \beta_1 t + \beta_2(t - t_1) \cdot \mathbb{I}(t > t_1) + \beta_3(t - t_2) \cdot \mathbb{I}(t > t_2) + \varepsilon$$

où :
— m(t) représente la valeur prédite (tendance)
— B0, B1, B2, B3 sont les coefficients du modèle
— t est le temps
— t1 et t2 sont les deux points de rupture
— I(t > t1) et I(t > t2) sont les indicateurs
— e représente les résidus

On retrouve dans cette sortie R les résultats de notre modèle :

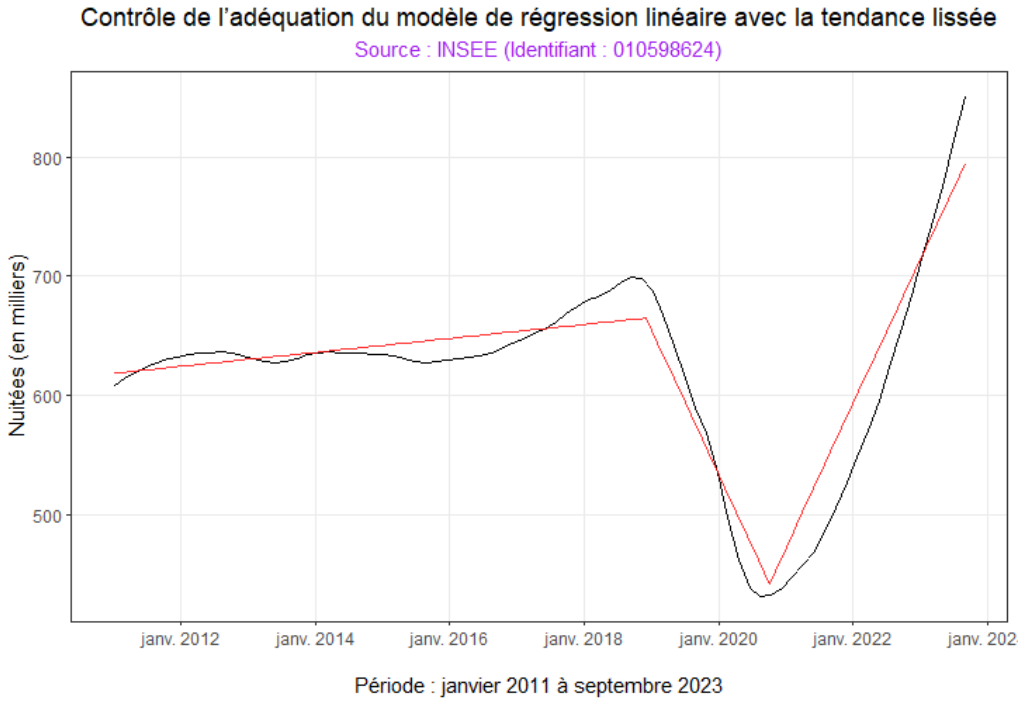
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 378.701800  17.735238   21.35   <2e-16 ***
time          0.016015   0.001073   14.93   <2e-16 ***
time1        -0.348855   0.005452  -63.99   <2e-16 ***
time2         0.663748   0.008241   80.55   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.606 on 149 degrees of freedom
Multiple R-squared:  0.9785,    Adjusted R-squared:  0.9781
F-statistic:  2261 on 3 and 149 DF,  p-value: < 2.2e-16
```

3.2 Vérifications d'adéquations

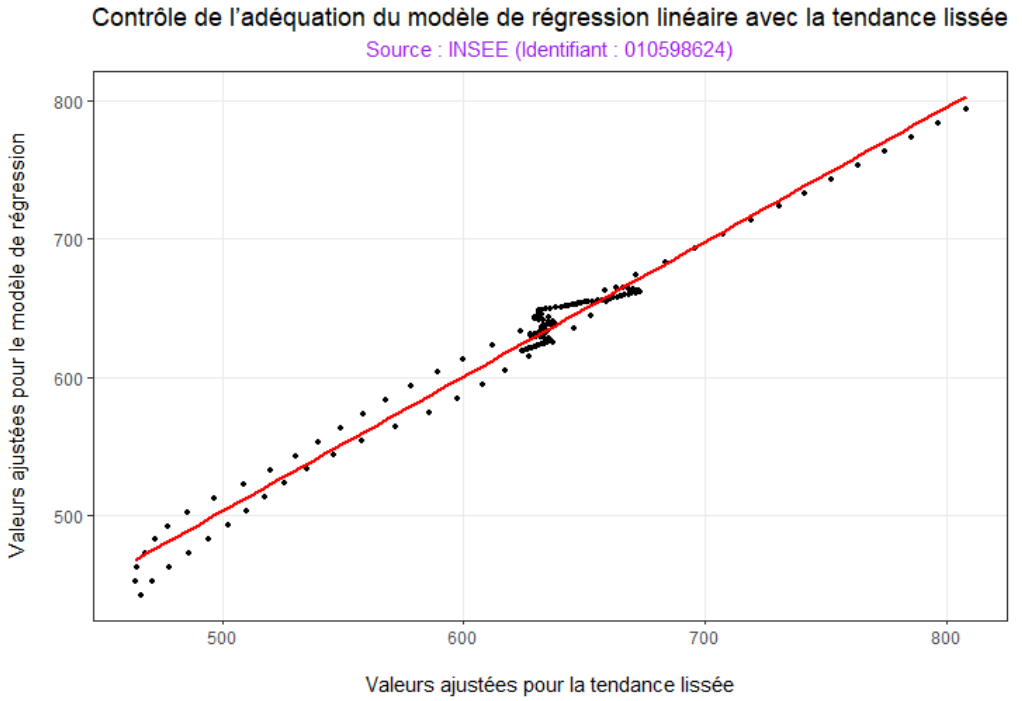
Adéquation du modèle de régression avec la tendance lissée

Nous allons maintenant vérifier l'adéquation de notre modèle de régression linéaire par morceaux avec la courbe de tendance lissée.

On remarque que la droite de régression linéaire est très semblable à la courbe de régression lissée ce qui indique un ajustement satisfaisant. On observe également grâce à ce graphique la présence des points de ruptures.



Pour contrôler davantage l'adéquation entre le modèle et la tendance lissée, nous avons créé le nuage de dispersion illustré ci-dessous. En examinant les valeurs ajustées de la tendance lissée par rapport aux valeurs ajustées du modèle de régression, nous constatons des divergences, ce qui suggère une relation. Dans l'ensemble, nos points sont répartis autour de la droite, indiquant ainsi un bon ajustement de notre modèle.



3.3 Modèle d'estimation de la composante saisonnière

Création du modèle de régression linéaire

A présent nous avons créé un second modèle afin d'estimer la composante saisonnière. Nous avons vu que nous avions un schéma additif, nous obtenons donc le modèle suivant :

$$Y_t = m(t) + S_t + \varepsilon_t$$

Où :
— St représente les variations saisonnières
— et représente les résidus
— m(t) représente la tendance

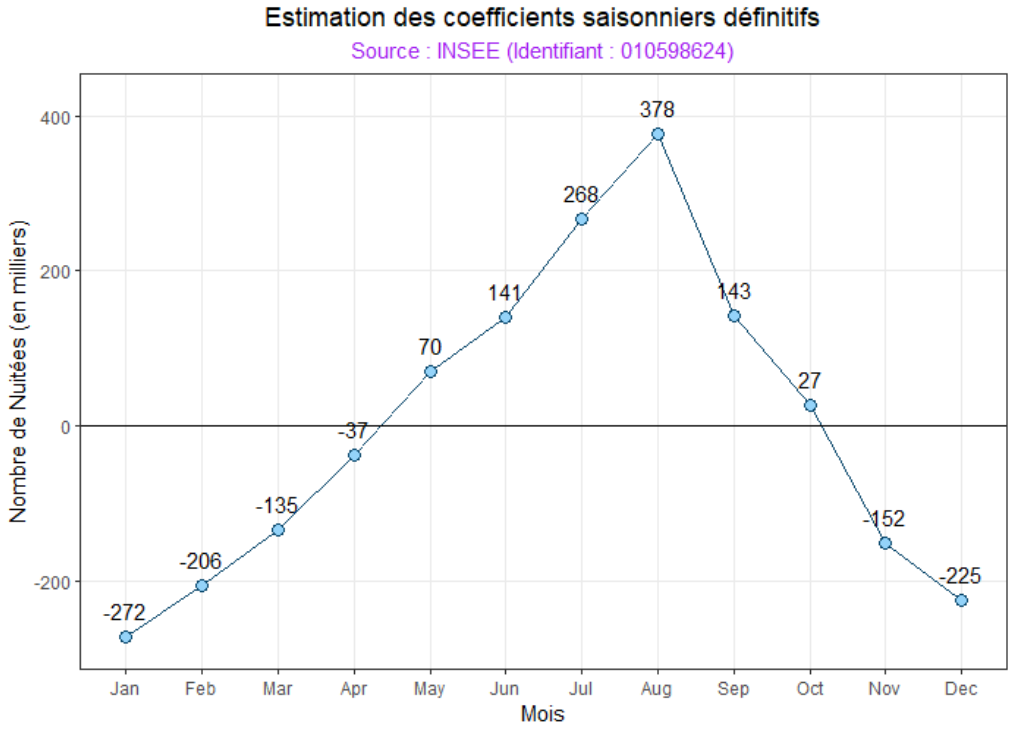
Ce modèle de régression linéaire explique 83,96 % de la variabilité de la série des variations saisonnières. Le mois d'août est le mois le plus favorable au nombre de nuitées dans les hôtels suivi des mois de juillet, septembre et juin, tandis que le mois de janvier correspond au mois le plus défavorable suivi des mois de décembre, février, novembre et mars. Pour le mois d'août, il faut s'attendre en moyenne à une augmentation de 373 000 nuitées dans l'hôtellerie indépendamment de la tendance. Pour le mois de janvier, il faut s'attendre à une diminution de 276 000 nuitées dans l'hôtellerie indépendamment de la tendance.

Sous R, nous obtenons ainsi ces résultats pour ce modèle :

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
Month_janv.  -276.44      25.07 -11.027  < 2e-16 ***
Month_févr.  -210.38      25.07  -8.392 4.54e-14 ***
Month_mars   -139.10      25.07  -5.549 1.38e-07 ***
Month_avr.    -41.80      25.07  -1.667  0.09764 .
Month_mai      65.61      25.07   2.617  0.00984 **
Month_juin    136.37      25.07   5.440 2.30e-07 ***
Month_juil.   263.16      25.07  10.497  < 2e-16 ***
Month_août    373.16      25.07  14.885  < 2e-16 ***
Month_sept.   138.15      25.07   5.511 1.65e-07 ***
Month_oct.     22.79      26.09   0.873  0.38400
Month_nov.   -156.75      26.09  -6.007 1.53e-08 ***
Month_déc.   -229.22      26.09  -8.785 4.81e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 90.39 on 141 degrees of freedom
Multiple R-squared:  0.8396,    Adjusted R-squared:  0.826
F-statistic: 61.51 on 12 and 141 DF,  p-value: < 2.2e-16
```

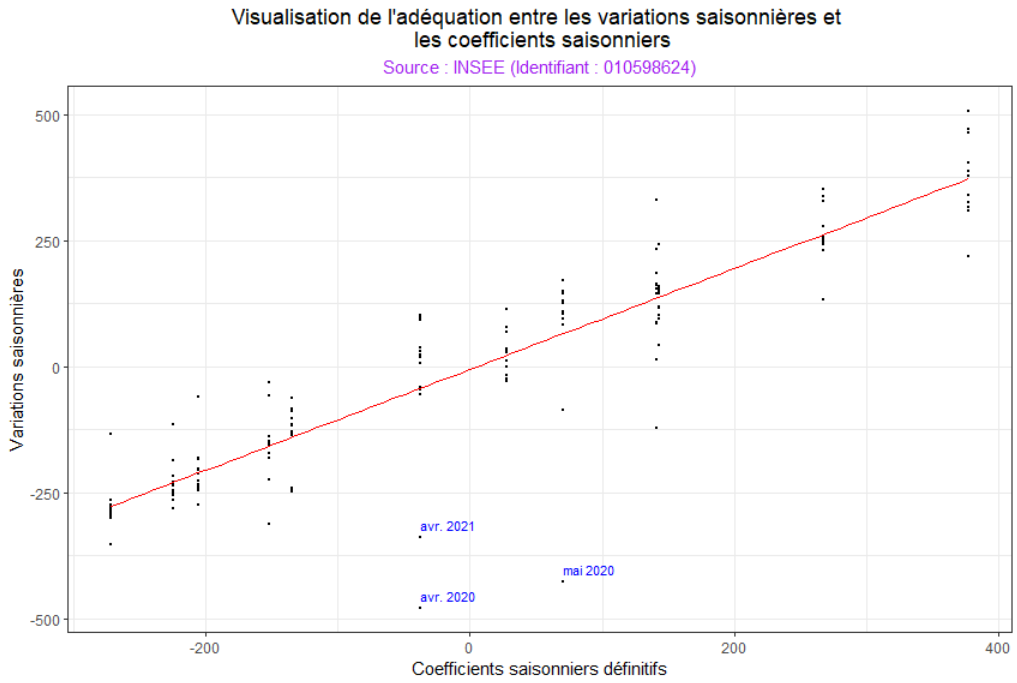
Représentation des coefficients saisonniers définitifs

L'illustration ci-dessous représente la répartition des coefficients saisonniers définitifs. On retrouve ainsi le mois le plus favorable au nombre de nuitées dans les hôtels, août, avec une augmentation de 378 000 nuitées par rapport à la tendance, c'est également le pic de notre graphique. À l'inverse, les mois de janvier et décembre sont les mois les plus défavorables avec respectivement une diminution de 225 000 et 272 000 nuitées par rapport à la tendance. On peut également noter que les mois les plus favorables sont les mois représentant les vacances d'été, celles-ci semblent attirer plus de clientèle dans les hôtels.

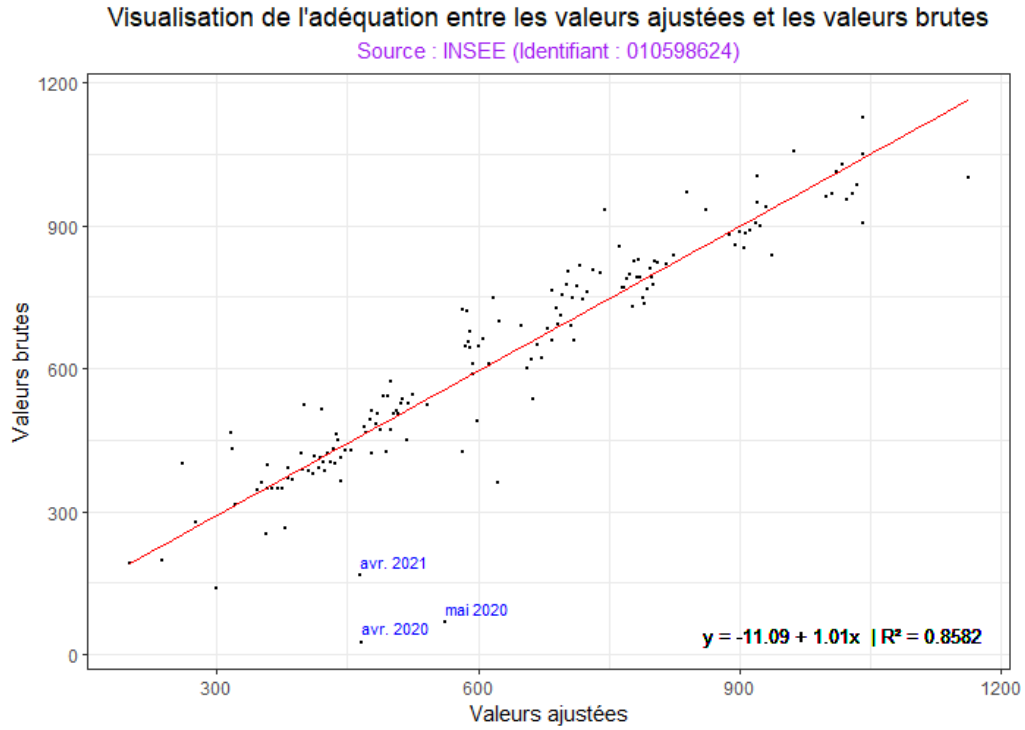


Vérification d'adéquation des variations et des coefficients saisonniers

Nous allons maintenant examiner la correspondance entre les fluctuations saisonnières et les coefficients saisonniers en utilisant le nuage de dispersion ci-dessous. On observe que les données varient lorsque la valeur d'une variable diffère, indiquant ainsi une association. De plus, nos valeurs sont disposées autour d'une droite. Il est cependant important de noter trois valeurs qui se démarquent des autres : avril 2021 et 2020, ainsi que mai 2020.



Nous allons maintenant visualiser la correspondance entre les valeurs ajustées et les valeurs brutes à l'aide du nuage de dispersion ci-dessous. On observe que les données varient lorsque la valeur d'une variable diffère, indiquant ainsi une association. Notre coefficient de détermination étant de 85,82 %, cela indique une association forte. De plus, nos valeurs sont disposées autour d'une droite. Trois valeurs se distinguent des autres valeurs, dont voici les dates associées : avril 2021 et 2020, ainsi que mai 2020.



4 Analyse des résidus

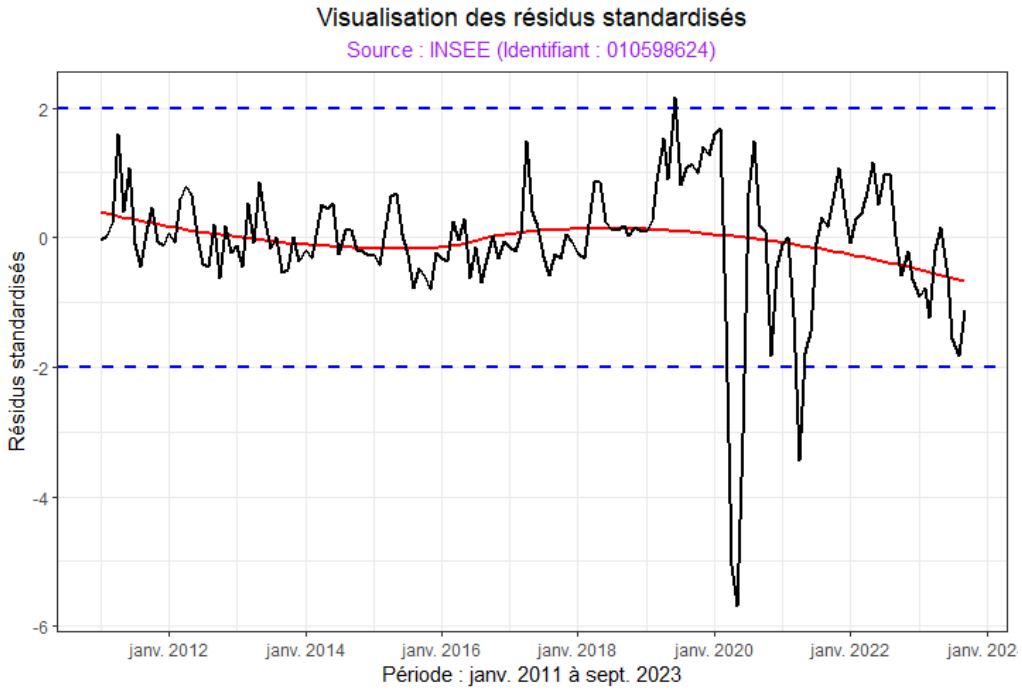
Dans cette phase, notre objectif est d'évaluer les résidus pour déterminer si des informations sont encore présentes dans les résidus. Si les résidus ne présentent pas un comportement de bruit blanc, cela indique la présence d'informations non prises en compte par le modèle. Pour qualifier les résidus de bruit blanc, trois critères doivent être respectés : ils doivent constituer un processus centré et homoscédastique, suivre une distribution gaussienne et présenter une absence de corrélation.

Pour ce faire voici les données que nous obtenons :

```
Rows: 153
Columns: 8
$ Date                  <fct> janv., févr., mars, avr., mai, juin, juil., août, sept., oc…
$ Varseason             <dbl> -274.53252, -203.02900, -115.47743, 102.02609, 105.54563, 2…
$ Année                 <chr> "2011", "2011", "2011", "2011", "2011", "2011", "2011", "20…
$ Coefficient           <dbl> -271.90578, -205.83734, -134.56380, -37.26459, 70.14656, 14…
$ Brutes                <int> 344, 416, 504, 722, 726, 855, 881, 961, 771, 691, 466, 388,…
$ Ajuste                <zoo> 346.6267, 413.1917, 484.9136, 582.7093, 690.6009, 761.8582,…
$ residual              <zoo> -2.626737, 2.808343, 19.086369, 139.290680, 35.399070, 93.1…
$ standardized_residual <zoo> -0.03017208, 0.03225811, 0.21923604, 1.59996582, 0.40661229…
```

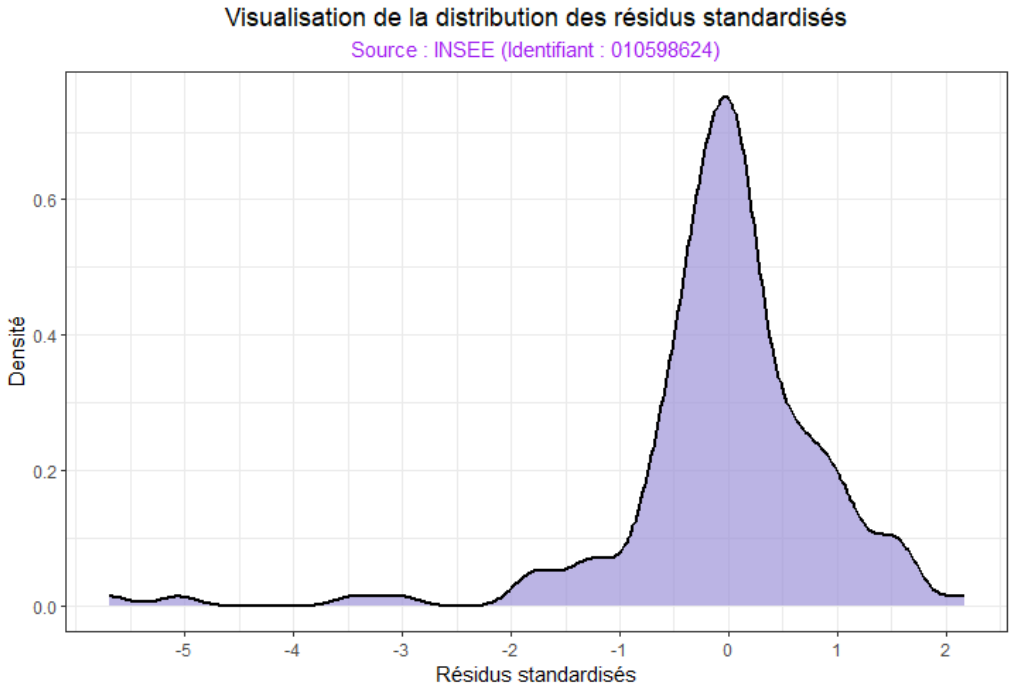
4.1 Visualisation de la série des résidus standardisés

Le graphique présenté ci-dessous cherche à évaluer si nos résidus suivent une loi gaussienne. Pour cela, à un niveau de signification de 5 %, il est nécessaire que 95 % des résidus standardisées se situent entre nos deux droites en pointillés bleus. De plus, la tendance lissée représentée en rouge, nous donne une indication de la répartition symétrique de nos valeurs, celle-ci doit s'approcher d'une ligne droite en 0 sous l'hypothèse d'un bruit blanc gaussien. Ainsi, on visualise que nos valeurs ne semblent pas tout à fait symétriques avec une tendance lissée qui fluctue légèrement. Elle témoigne également d'une faible variabilité des valeurs démontrant ainsi une homoscédasticité. L'ensemble des valeurs semblent situées dans notre bande de confiance. On note néanmoins deux points atypiques montrant deux pics fortement négatifs en juin 2020 et avril 2021.

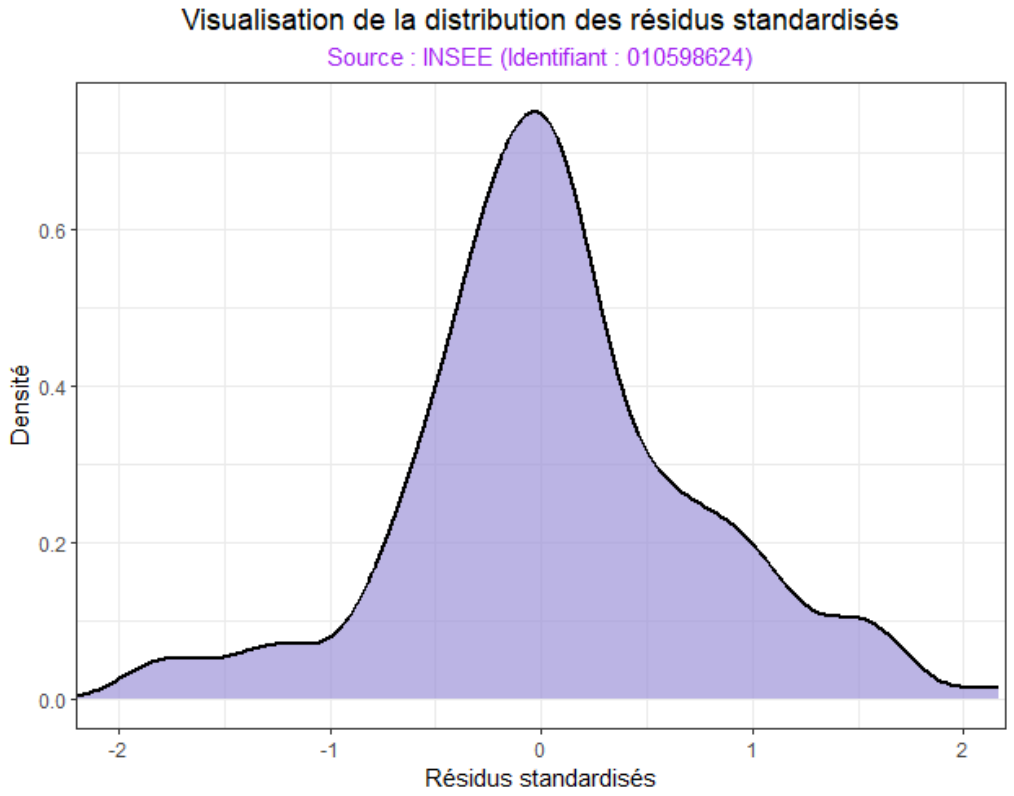


4.2 Visualisation de la densité des résidus standardisés

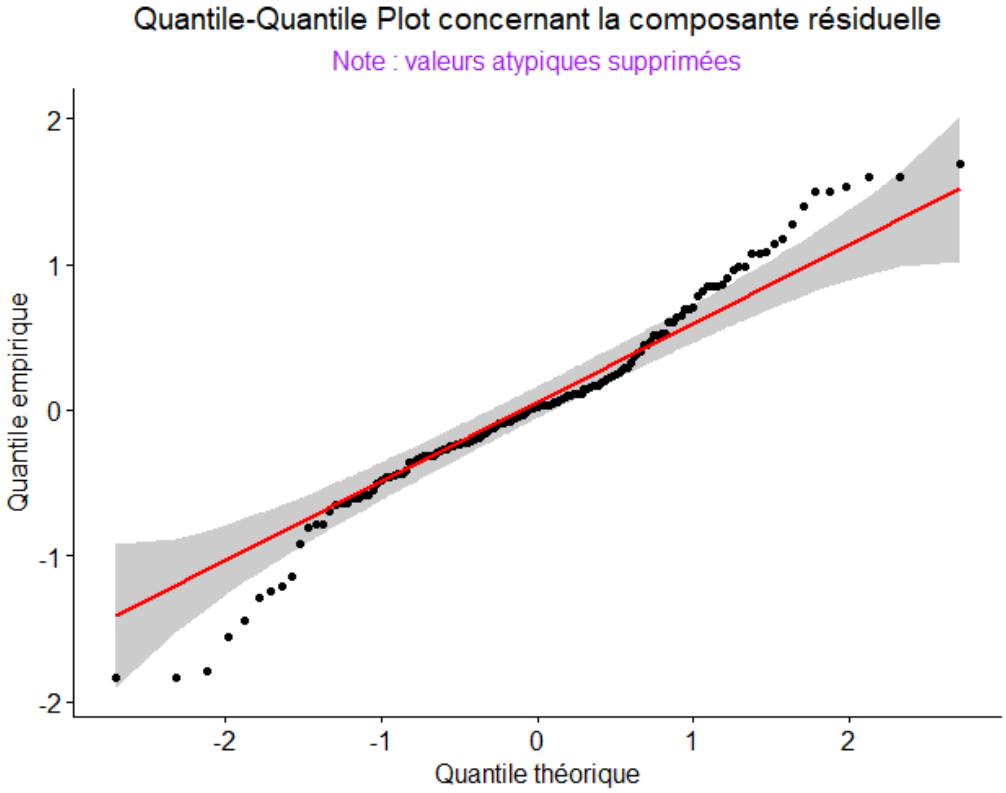
Afin de visualiser plus clairement si nos données semblent suivre une loi gaussienne, nous avons dressé la courbe de densité des résidus standardisés. La figure ci-dessous met en évidence une distribution unimodale avec un aplatissement à gauche. Cet aplatissement est dû à la présence de valeurs fortement atypiques.



Afin de visualiser plus clairement la distribution des résidus standardisés, nous avons dressé la courbe de densité sans les valeurs atypiques observées précédemment, la figure ci-après présente la courbe de densité des résidus standardisés en excluant les données atypiques. On observe un aplatissement à droite de la courbe. La distribution bien qu'unimodale, celle-ci ne semble pas symétrique. Cela nous laisse penser que les résidus standardisés ne suivent pas une loi gaussienne.



Un autre indicateur permettant d'évaluer si nos données suivent ou non une loi gaussienne est la mise en œuvre d'un quantile-quantile plot. Celui-ci, compare les quantiles empiriques des résidus standardisés à ceux des quantiles théoriques d'une distribution normale. Ainsi, on observe via la figure représentée ci-dessous la présence de points en dehors de la zone de confiance. Cela vient renforcer l'hypothèse faite précédemment, les résidus standardisés ne semblent pas suivre une loi gaussienne.



4.3 Test de Shapiro-Wilk

Pour vérifier si la distribution n'est pas normale comme nous l'avons supposé précédemment, nous avons effectué le test de Shapiro-Wilk. Le niveau de signification est de 5 %. Nous avons les deux hypothèses suivantes :

Hypothèse nulle : *la distribution suit une loi gaussienne*
VS
Hypothèse alternative : *la distribution ne suit pas une loi gaussienne*
Nous obtenons la sortie suivante :

```
        Shapiro-Wilk normality test

data:  data
W = 0.97333, p-value = 0.00555
```

Ce test nous rapporte une p-value égale à 0,5 %. La p-value étant inférieure à 5 %, nous devons rejeter l'hypothèse nulle (la distribution suit une loi gaussienne) au profit de l'hypothèse alternative (la distribution ne suit pas une loi gaussienne). Au vu de ces résultats, nous pouvons dire que la distribution des résidus standardisés ne suit pas une loi gaussienne.

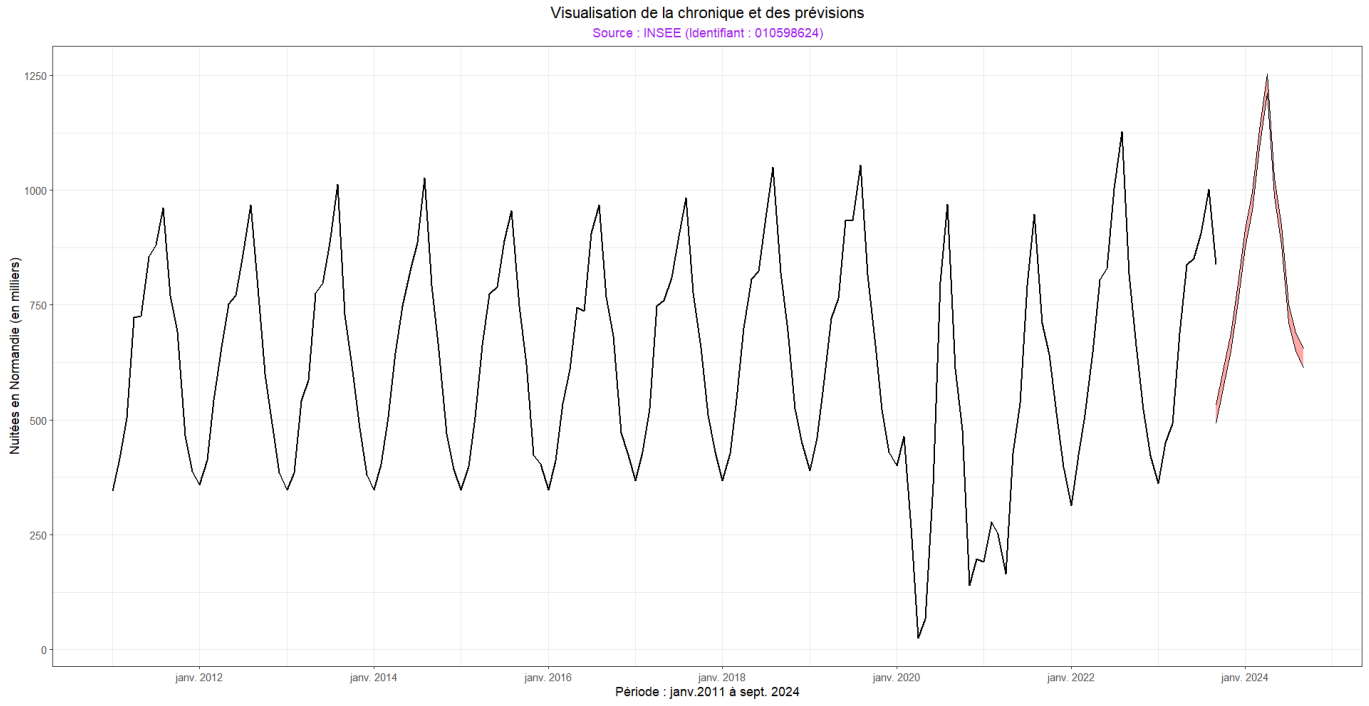
5 Prévision des données

Dans cette dernière partie, nous cherchons à faire une prédiction des données jusqu'en septembre 2024. Nous voulons avoir également un intervalle de confiance de cette prédiction. Pour ce faire, nous avons calculé les données présentes dans la sortie ci-dessous :

	Date	Value.fit	Value.lwr	Value.upr
1	2023-09-01	511.8766	492.0228	531.7305
2	2023-10-01	587.8723	567.9544	607.7902
3	2023-11-01	669.4040	649.4169	689.3912
4	2023-12-01	776.6305	756.5733	796.6877
5	2024-01-01	894.2998	874.1672	914.4324
6	2024-02-01	975.3187	955.1077	995.5298
7	2024-03-01	1111.7108	1091.4237	1131.9980
8	2024-04-01	1231.9605	1211.5891	1252.3320
9	2024-05-01	1006.8793	986.4235	1027.3351
10	2024-06-01	901.7765	881.2306	922.3223
11	2024-07-01	732.1680	711.5323	752.8038
12	2024-08-01	669.9562	649.2248	690.6876
13	2024-09-01	632.9892	612.1593	653.8190

5.1 Visualisation de la prévision

Le graphique ci-dessous illustre la chronique représentant l'évolution du nombre de nuitées dans l'hôtellerie en Normandie. Nous avons ajouté à ce graphique l'intervalle de prédiction représentée en rouge qui va jusqu'en septembre 2024. On remarque que le nombre de nuitées continuera d'augmenter jusqu'à même dépasser les valeurs avant le premier point de rupture représentant le début du confinement suite au COVID-19. Ces prédictions sont valables que si la situation ne change pas. Les facteurs qui peuvent changer cette évolution ne sont pas pris en compte dans les calculs de la prédiction.



6 Conclusion

En résumé, l'objectif de ce projet était de réaliser une prédiction sur le nombre de nuitées dans l'hôtellerie en Normandie pour l'année à venir. Pour ce faire, nous avons développé des modèles à la suite d'une analyse exploratoire. Cette analyse nous a permis de nous familiariser avec les données, d'identifier les points atypiques potentiels et les points de rupture, formant ainsi notre modèle de régression par morceaux pour estimer la tendance sur différentes périodes. Nous avons déterminé trois périodes distinctes : la première avant décembre 2018, la deuxième de décembre 2018 à octobre 2020, et la troisième d'octobre 2020 à septembre 2023.

De manière générale, la tendance connaît une augmentation moyenne de 16 nuitées par mois, indépendamment de la tendance observée lors de la première période. Cette augmentation est suivie d'une chute abrupte, avec une diminution moyenne de 349 nuitées par mois entre décembre 2018 et octobre 2020. À partir d'octobre 2020, une reprise significative est constatée, avec une augmentation moyenne de la tendance de 664 nuitées par mois.

En ce qui concerne la composante saisonnière de notre série temporelle, nous observons que la période estivale, notamment le mois d'août, est la plus favorable, tandis que la période hivernale, de décembre à février, est la moins favorable, avec le mois de janvier étant particulièrement défavorable.

En ce qui concerne la qualité des modèles, nous constatons que de l'information subsiste dans nos résidus standardisés, comme le révèle la création du corrélogramme. En effet, ces résidus standardisés ne se comportent pas comme un bruit blanc, présentant une corrélation entre l'instant t et t + 1. De plus, leur distribution ne suit pas une loi gaussienne.

Les prévisions pour l'année 2024 indiquent une augmentation significative du nombre de nuitées, atteignant un pic de 1 250 000 nuitées pendant la période estivale. Avant la pandémie de COVID-19, le chiffre était d'environ 1 000 000 de nuitées dans l'hôtellerie en Normandie.